Estimation and Test for Two Pareto Populations with Partially Missing Data

Daozhong Luo

School of Mathematical Sciences, Huaqiao University
Quanzhou, Fujian 362021, P. R. China
ldzblue@163.com

Copyright © 2013 Daozhong Luo. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

This paper deals with the estimation of a parameter for the Pareto distribution, and proves strong consistency and asymptotic normality of the estimators. Moreover, statistics on testing the equality of two populations and its limit distribution are given.

Keywords: Pareto distribution; Missing data; Limit distribution

1 Introduction

Pareto distribution is used to describe issues such as personal income, risk of insurance, commercial invalidation, the survival time of patients after a pharmacological process, etc. Cases of missing data often take place when we apply statistics to solve corresponding problems. There are a lot of papers in literature discussing statistics assumption of Poisson distribution and two exponential populations with missing data (relevant results can see[1],[2] and[3]). In this paper, we deal with maximum likelihood estimation of parameter and hypothesis test of two equal parameter populations when neither of two Pareto distribution populations is utterly under control of observers. The maximum likelihood estimation of parameter is given, and strong consistency and asymptotic normality are proved. Furthermore, statistic on testing the equality of two populations and its limit distribution are given.
2 Hypothesis and maximum likelihood estimation

For two Pareto distribution populations, the density function is

\[ f_i(x; \theta_i; \alpha_i) = \begin{cases} \frac{\theta_i \alpha_i^\theta_i}{x^{\alpha_i+1}} & x > \alpha_i \\ 0 & x \leq \alpha_i \end{cases} \quad (i = 1, 2), \]

where \( \alpha_i > 0 \) \((i = 1, 2)\) is a given parameter, and \( \theta_i > 1 \) \((i = 1, 2)\) is an unknown one. Now we make \( n \) times observation on the two populations respectively. The samples are denoted as \((X_1, X_2, \cdots, X_n), (Y_1, Y_2, \cdots, Y_n)\), and missing data is found in both observations. For the first population, \( X_j \) is missing with probability \( 1 - p_1 \), so the actual observed data is \((X_j, \delta_j) \quad (j = 1, 2, \cdots, n)\), where \((X_1, X_2, \cdots, X_n)\) and \((\delta_1, \delta_2, \cdots, \delta_n)\) are independent, and \( \delta_1, \delta_2, \cdots, \delta_n \) is identically independent distribution with \( \delta_i \sim b(1, p_1) \). In addition, \( \delta_j = 1 \) denotes that \( X_j \) is observed; while \( \delta_j = 0 \), \( X_j \) means oppositely. If we rewrite the actual \( n_i \) observed value (when \( n_i = 0 \), sampling should be carried on) obtained in the first population as \((Z_{i1}, Z_{i2}, \cdots, Z_{in_i})\), clearly,

\[ n_i = \sum_{j=1}^n \delta_j, \quad \sum_{j=1}^{n_i} Z_{ij} = \sum_{j=1}^n X_j \delta_j. \]

Similarly, considering the second population, \( Y_j \) is missing with probability \( 1 - p_2 \), the actual observed data is \((Y_j, \eta_j) \quad (j = 1, 2, \cdots, n)\), where \((Y_1, Y_2, \cdots, Y_n)\) and \((\eta_1, \eta_2, \cdots, \eta_n)\) are independent, and \( \eta_1, \eta_2, \cdots, \eta_n \) is identically independent distribution with \( \eta_i \sim b(1, p_2) \). In addition, \( \eta_j = 1 \) denotes that \( Y_j \) is observed; while \( \eta_j = 0 \),
Estimation and test for two Pareto populations

$Y_j$ denotes that there is no observed value obtained. Similarly, if we rewrite the actual $n_2$ observed value (when $n_2 = 0$, sampling should be carried on) obtained in the second population as $Z_{21}, Z_{22}, \ldots, Z_{2n_2}$, Thus,

$$n_2 = \sum_{j=1}^{n_1} \eta_j, \sum_{j=1}^{n_2} Z_{2j} = \sum_{j=1}^{n_1} Y_j \eta_j.$$

To discuss maximum likelihood estimation of $\theta_1$, We get the likelihood function

$$L(\theta_1) = \prod_{j=1}^{n_1} f_1(Z_{1j}; \theta_1) = \prod_{j=1}^{n_1} \frac{\theta_1 \alpha_1^{\theta_1 - 1}}{Z_{1j}^{\theta_1}}.$$

Thus, the maximum likelihood estimation of $\theta_1$ is easily to be calculated as followed,

$$\hat{\theta}_1 = \left( \frac{1}{n_1} \sum_{j=1}^{n_1} \ln Z_{1j} - \ln \alpha_1 \right)^{-1}.$$

Using the similar way, the maximum likelihood estimation of $\theta_2$ is obtained as followed,

$$\hat{\theta}_2 = \left( \frac{1}{n_2} \sum_{j=1}^{n_2} \ln Z_{2j} - \ln \alpha_2 \right)^{-1}.$$

If we consider the hypothesis: $H_0: \theta_1 = \theta_2 = \theta$, where $\theta$ is unknown, the observation likelihood function of $\theta$ is

$$L(\theta) = \prod_{j=1}^{n_1} \frac{\theta \alpha_1^{\theta}}{Z_{1j}^{\theta+1}} \prod_{j=1}^{n_2} \frac{\theta \alpha_2^{\theta}}{Z_{2j}^{\theta+1}}.$$

Thus we calculate the maximum likelihood estimation of $\theta$ as followed,

$$\hat{\theta} = \left( \frac{1}{n_1 + n_2} \left( \sum_{j=1}^{n_1} \ln Z_{1j} + \sum_{j=1}^{n_2} \ln Z_{2j} \right) - \frac{n_1}{n_1 + n_2} \ln \alpha_1 - \frac{n_2}{n_1 + n_2} \ln \alpha_2 \right)^{-1}.$$
3 Asymptotic properties of maximum likelihood estimation

Lemma 1 Suppose \( \{Z_n\} (n = 1, 2, \cdots) \) is a random variable sequence, and \( Z_n \longrightarrow c \) a.s., where \( c \) is a constant and function \( g(\bullet) \) is continuous at \( c \), then \( g(Z_n) \longrightarrow g(c) \) a.s.

Lemma 2 Suppose \( \{a_n\} (n = 1, 2, \cdots) \) is a sequence tending to \( +\infty \), \( b \) is a constant, and \( \{Z_n\} (n = 1, 2, \cdots) \) is a random variable sequence, and \( a_n(Z_n - b) \overset{d}{\longrightarrow} Z \). Suppose \( g(\bullet) \) is a differentiable function, and \( g'(\bullet) \) is continuous at \( b \), then \( a_n(g(Z_n) - g(b)) \overset{d}{\longrightarrow} g'(b)Z \).

The strong consistency and asymptotic normality of \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) are respectively given in the following theorems.

Theorem 1 Consistently with the previous denotation,
\[ \hat{\theta}_i \longrightarrow \theta_i \text{ a.s. } (i = 1, 2). \]

Proof It is enough to show that \( \hat{\theta}_i \longrightarrow \theta_i \) a.s. Let \( H_i = \ln X_i, i = 1, 2, \cdots, n \).

It follows that \( \frac{1}{n} \sum_{j=1}^{n} \ln Z_{ij} = \sum_{j=1}^{n} H_j \delta_j \) and \( EH_i = \ln \alpha_i + \frac{1}{\theta_i^2}, \text{DH}_i = \frac{2}{\theta_i^2} \). Note that \( \{H_j \delta_j: 1 \leq j \leq n\} \) is i.i.d., By employing the law of large numbers, we obtain
\[ \frac{1}{n} \sum_{j=1}^{n} H_j \delta_j \quad \text{EH}_i \delta_i = \text{EH}_i E\delta_i = p_i \left( \ln \alpha_i + \frac{1}{\theta_i^2} \right) \text{ a.s. } , \]
and \[ \frac{n_i}{n} = \frac{1}{n} \sum_{j=1}^{n} \delta_j \quad \text{p}_i \text{ a.s. } . \]

It follows from Slutsky Theorem that
Estimation and test for two Pareto populations

\[
\frac{1}{\theta_i} = n \frac{1}{n_i} \sum_{j=1}^{n_i} H_j \delta_j - \ln \alpha_i \to \frac{1}{p_i} \ln \alpha_i \left( \ln \alpha_i + \frac{1}{\theta_i} \right) - \ln \alpha_i = \frac{1}{\theta_i} \text{ a.s.}
\]

Let \( g(t) = \frac{1}{t} \). Clearly \( g(t) \) is continuous at \( \frac{1}{\theta_i} \). By lemma 1, we show that

\[
\hat{\theta}_i = g \left( \frac{1}{\theta_i} \right) \to g \left( \frac{1}{\theta_i} \right) = \theta_i \text{ a.s. } \]

\[\Box\]

**Theorem 2** \( \sqrt{n} \left( \hat{\theta}_i - \theta \right) \xrightarrow{\text{L}} N \left( 0, \frac{2\theta_i^2}{p_i} \right) \) (\(i = 1, 2\))

**Proof** It is only need to prove when \( i = 1 \).

\[
\sqrt{n} \left( \frac{1}{\theta_i} - \frac{1}{\theta_i} \right) = \sqrt{n} \left( \frac{1}{n_i} \sum_{j=1}^{n_i} \ln Z_{ij} - \ln \alpha_i - \frac{1}{\theta_i} \right) = \sqrt{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \left( H_j - \ln \alpha_i - \frac{1}{\theta_i} \right) \delta_j
\]

\[
= \frac{n}{n_i} \sum_{j=1}^{n_i} \left( H_j - \ln \alpha_i - \frac{1}{\theta_i} \right) \delta_j,
\]

Note \( E \left( H_j - \ln \alpha_i - \frac{1}{\theta_i} \right) \delta_i = E \left( H_j - \ln \alpha_i - \frac{1}{\theta_i} \right) E \delta_i = 0 \), and

\[
D \left( H_j - \ln \alpha_i - \frac{1}{\theta_i} \right) \delta_i = E \left[ \left( H_j - \ln \alpha_i - \frac{1}{\theta_i} \right)^2 \delta_i \right] = E \left( H_j - \ln \alpha_i - \frac{1}{\theta_i} \right)^2 E \delta_i = \frac{2p_i}{\theta_i^2}.
\]

By the Central limit theorem, we obtain

\[
\sqrt{n} \frac{2p_i}{\theta_i^2} \xrightarrow{\text{L}} N \left( 0, 1 \right),
\]

i.e.

\[
\sqrt{n} \sum_{j=1}^{n_i} \left( H_j - \ln \alpha_i - \frac{1}{\theta_i} \right) \delta_j \xrightarrow{\text{L}} N \left( 0, \frac{2p_i}{\theta_i^2} \right).
\]

Obviously, \( \frac{n_i}{n} \to p_i \) a.s. implying \( \frac{n_i}{n} \xrightarrow{p} p_i \).
\[ n \frac{1}{n_1 \sqrt{n}} \sum_{j=1}^{n} \left( H_j - \ln \alpha_i - \frac{1}{\hat{\theta}_i} \right) \delta_j \xrightarrow{\mathcal{L}} N \left( 0, \frac{2}{p_i \hat{\theta}_i^2} \right). \]

Let \( g(t) = \frac{1}{t} \), then
\[ \sqrt{n} \left( \hat{\theta}_i - \theta \right) = \sqrt{n} \left( g \left( \frac{1}{\hat{\theta}_i} \right) - g \left( \frac{1}{\theta} \right) \right) \xrightarrow{\mathcal{L}} -\theta^2 N \left( 0, \frac{2p_i}{\theta^2} \right), \]
i.e.
\[ \sqrt{n} \left( \hat{\theta}_i - \theta \right) \xrightarrow{\mathcal{L}} N \left( 0, \frac{2\theta^2}{p_i} \right). \]

**Corollary**  
With the hypothesis \( H_0 : \theta_1 = \theta_2 = \theta, \hat{\theta} \rightarrow \theta \) a.s.,

and
\[ \sqrt{n} \left( \hat{\theta} - \theta \right) \xrightarrow{\mathcal{L}} N \left( 0, \frac{2\theta^2}{p_1 + p_2} \right). \]

### 4 Test of two populations with equal parameter and asymptotic confidence interval of parameter difference of two populations.

By theorem1 and theorem2, it is easy to show

**Theorem 3**
\[ \sqrt{n} \left[ \left( \hat{\theta}_1 - \hat{\theta}_2 \right) - \left( \theta_1 - \theta_2 \right) \right] \xrightarrow{\mathcal{L}} N (0, 1). \]

Especially, if \( H_0 \) is valid, then
\[ \sqrt{n} \left( \hat{\theta}_1 - \hat{\theta}_2 \right) \xrightarrow{\mathcal{L}} N (0, 1). \]Thus, under the confidence level \( \alpha \) (\( 0 < \alpha < 1 \)), asymptotic confidence interval of \( \theta_1 - \theta_2 \) is
\[ \left( \hat{\theta}_1 - \hat{\theta}_2 - U \frac{2\theta^2}{p_1} + U \frac{2\theta^2}{p_2} \right), \hat{\theta}_1 - \hat{\theta}_2 + U \frac{2\theta^2}{p_1} \]

**Remark 1** If \( n_1 = 0 \) or \( n_2 = 0 \), then there is no valid observed data. When the parameter is thought to be 0, either estimation or test is invalidated. In practice, if
\( n_1 = 0 \) or \( n_2 = 0 \), instead of taking the assumed \( H_0 \), we should continue to take samples until valid observation is obtained.

**Remark 2** If the samples of the two populations are not equal, we still can have the similar discussion.

**References**


**Received: January 15, 2013**